Machine Learning External Lab Exam

Set 2 – Lab Exam

PART A – Image Classification using EMNIST Letters (Total: 22 marks)

Dataset Description:
You are given a CSV file named set_2_dataset_1.csv, which contains data from the EMNIST Letters dataset. This dataset includes 28x28 grayscale images of handwritten English letters (A–Z).

The first column contains labels ranging from 1 to 26, where: 1 = A, 2 = B … 26 = Z

The remaining 784 columns contain pixel values of each image, in row-major order.

1. Load the dataset set_2_dataset_1.csv. (No marks)
2. Print the shape of the dataset. (1 mark)
3. Print the unique classes and their counts. (2 marks)
4. Display one image by reshaping it to 28 x 28 and print its label. (2 marks)
5. Display 6 images using subplot with their labels. (2 marks)
6. Prepare features (X) and labels (y), split the dataset (15% test), and print the shape of training and testing sets. (3 marks)
7. Train a suitable classification model. (2 marks)
8. Predict on the test data. (1 mark)
9. Print the confusion matrix. (1 mark)
10. Print the classification report. (1 mark)
11. Display 6 test images with actual and predicted labels. (2 marks)
12. Apply GridSearchCV on one parameter. (3 marks)
13. Print the best score and best parameter from GridSearchCV. (2 marks)

PART B – Regression using Concrete Strength Dataset (Total: 18 marks)

Dataset Description:
You are provided with set_2_dataset_2.csv, which contains data related to concrete mixtures. The dataset includes features such as: Cement amount, Fly ash, Blast furnace slag, Water, Superplasticizer, Coarse and fine aggregates, Age (in days)

The target variable is compressive *strength* of the concrete, measured in MPa(megapascals).

14. Load the dataset set_2_dataset_2.csv. (No marks)
15. Print the summary statistics. (1 mark)
16. Print the number of missing values in each column. (1 mark)
17. Drop rows with missing values in output column (Strength column). (1 mark)
18. Impute the missing values using pandas built-in functions. (2 marks)
19. Plot a histogram of the target variable ‘Strength’ to visualize its distribution. (2 marks).
20. Plot scatter plot of Water and Cement vs Strength in two subplots.. (2 marks)

21. Prepare features (X) and label (y), split the dataset (15% test). (2 marks)
22. Train a regression model. (2 marks)
23. Print the R^2 value on the test data. (2 marks)
24. Predict the output for a sample test data point. (2 mark)
25. Predict the regression coefficients. (1 mark)

Set 2 – Viva Questions (20 Marks)

1. In Naive Bayes classification, we assume that the features are _____ given the class label.
2. The sigmoid function maps input values into the range of _____.
3. In Decision Tree, the measure of uncertainty or impurity is called _____.
4. _____ Regression is used for predicting binary outcomes.
5. In K-Means Clustering, each cluster is associated with a _____.
6. The hyperplane in SVM is defined as the equation _____. (Write the equation)
7. The purpose of the elbow method in K-Means clustering is to find the _____.
8. The dot product in SVM is used to compute the _____ from the hyperplane.
9. One-hot encoding is a method used to convert _____ data into numerical format.
10. In Logistic Regression, the decision boundary is typically set at a probability threshold of _____.
11. The most common distance metric used in KNN is the _____ distance.
12. The confusion matrix is used to evaluate the performance of a _____ model.
13. In data preprocessing, missing numerical values can be replaced using _____ imputation.
14. In supervised learning, the data contains both input features and _____ .
15. Entropy reaches its maximum when the dataset is _____ .
16. If a point lies exactly on the SVM hyperplane, its margin is equal to _____.
17. When comparing models, the metric _____ is preferred for regression tasks.
18. Data preprocessing includes steps like dealing with missing values, scaling, and _____.
19. In feature scaling, _____ normalization brings features into a 0-1 range.
20. A neuron triggers an output signal only if the combined input exceeds a certain _____.